(54) Title: DRUG DISCOVERY USING GENE EXPRESSION PROFILING

(57) Abstract: The invention relates to a method of determining a candidate condition that is treatable with a chemical composition by (1) providing a relational database comprising phenotype data values and gene expression profile data values, wherein each of the gene expression profile data values is related to at least one phenotype data value, and each of the phenotype data values is associated with a condition in an individual; (2) contacting a cell with the chemical composition; (3) obtaining a test gene expression profile of the cell after the contacting; (4) querying the relational database with the test gene expression profile to obtain a test phenotype; and (5) determining the candidate condition associated with the test phenotype.

DRUG DISCOVERY USING GENE EXPRESSION PROFILING

5

Background of the Invention

DNA chips allow relatively quick and easy generation of a gene expression profile in a particular cell or tissue. See, e.g., Southern et al., Trends Genet. 12:110-115, 1996; and Ginot,
10 Human Mutation 10:1-10, 1997. Additional aspects of DNA microarrays and drug development are discussed in De SaiZieu et al., Nature Biotechnology 16:45-48, 1998; Heller et al., Proc. Natl. Acad. Sci. USA 94:2150-2155, 1997; Lockhart, Nature Biotechnology 14:1675-1680, 1996; Fields et al., Pro. Natl. Acad.
15 Sci. USA 96:8825-8826, 1999; Lander, Nature Genetics S21:3-4, 1999; and Bowtell, Nature Genetics S21:25-32, 1999.

Summary of the Invention

The invention relates to new methods for finding
20 compositions or compounds that can be used to treat a condition or disease in an individual. These new methods do not require expensive or time-consuming biological assays but instead rely on the generation and analysis of gene expression profiles of different cells under various conditions. DNA chips help
25 facilitate these methods.

Accordingly, the invention features a method of determining a candidate condition (including a disease such as cancer, asthma, osteoporosis, Alzheimer's Disease, and diabetes) that is treatable with a chemical composition by (1) providing a
30 relational database having phenotype data values and gene expression profile data values, each of the gene expression profile data values relating to at least one phenotype data value, and each of the phenotype data values associated with a condition in an individual; (2) contacting a cell with the
35 chemical composition; (3) obtaining a test gene expression profile of the cell after the contacting; (3) querying the relational database with (or using) the test gene expression profile to obtain a test phenotype; and (4) determining the

candidate condition associated with the test phenotype.

In general gene expression profiles of a particular
cellular or cell-derived sample can be generated using a
combination of standard techniques, such as sequential or
5   parallel Northern blotting, differential display technologies, or
nucleic acid microarray technology, which will be discussed
below. A phenotype can include any cellular characteristic such
as metastatic, apoptotic, necrotic, or normal.

The database can further include cell identity data
10 values, each of the cell identity data values being related to at
least one of the phenotype data values and to at least one of the
gene expression profile data values.

The test gene expression profile can be obtained by a
process including (1) isolating mRNA from the cell; (2) producing
15 cDNA from the mRNA; (3) hybridizing the cDNA to an array of
nucleic acid elements (e.g., on a glass or silicon-based
support), each element corresponding to a gene; and (4) measuring
the extent to which cDNA has bound to each element, thereby
determining the test gene expression profile. To facilitate the
20 measuring step, the cDNA can be labeled during or after cDNA
synthesis with a label such as a radionuclide, fluorescent
molecule, luminescent molecule, or a chromogenic molecule such as
an enzyme. The composition in the above method can contain a
pure bioactive compound or mixture of a number of pure compounds,
25 or a crude extract including a plant extract or an extract of an
animal tissue.

The invention also includes a computer system having (1)
a memory storing a database including phenotype data values and
gene expression profile data values, each of the gene expression
30 profile data values relating to at least one phenotype data
value, and each of the phenotype data values associated with a
condition in an individual; (2) an input device configured to
provide a test gene expression profile obtained from a cell after
contacting the cell with a composition; and (3) a processor
35 configured by a program to query the database using the test gene
expression profile to obtain a test phenotype. The computer

system can further include an output device for conveying the
·test phenotype or the condition associated with the test
phenotype. In addition, the database can further include cell
identity data values, each of the cell identity data values being
5 related to at least one of the phenotype data values and to at
least one of the gene expression profile data values.

The invention also features a computer-readable medium
having a program adapted to configure a machine to query a
database with (using) a test gene expression profile to obtain a
10 test phenotype, the database including phenotype data values and
gene expression profile data values, each of the gene expression
profile data values relating to at least one phenotype data
value, and each of the phenotype data values associated with a
condition in an individual. The database can further include
15 cell identity data values, each of the cell identity data values
being related to at least one of the phenotype data values and to
at least one of the gene expression profile data values.

Databases useful in the invention can further include
condition data values associated with the condition in the
20 individual.

Also featured in the invention is a method of identifying
a candidate mixture (e.g., a plant, fungal, bacterial, or animal
tissue extract) for treating a condition in an individual, by (1)
providing a first gene expression profile of a conditioned cell,
25 the conditioned cell exhibiting a phenotype that can be
correlated with the condition in the individual; (2) contacting
the conditioned cell with the mixture having a plurality of
(e.g., at least 10 or 100) different types of bioactive
molecules; (3) determining a second gene expression profile of
30 the conditioned cell after the contacting; and (4) comparing
(e.g., by using a computer) the second gene expression profile
with the first gene expression profile. A change in the first
gene expression profile relative to the second gene expression
profile indicates that the mixture is a candidate composition for
35 treating the condition in the individual. A bioactive molecule
is a molecule that elicits a biochemical or cellular response in

- 3 -

system can further include an output device for conveying the test phenotype or the condition associated with the test phenotype. In addition, the database can further include cell identity data values, each of the cell identity data values being related to at least one of the phenotype data values and to at least one of the gene expression profile data values.

The invention also features a computer-readable medium having a program adapted to configure a machine to query a database with (using) a test gene expression profile to obtain a test phenotype, the database including phenotype data values and gene expression profile data values, each of the gene expression profile data values relating to at least one phenotype data value, and each of the phenotype data values associated with a condition in an individual. The database can further include cell identity data values, each of the cell identity data values being related to at least one of the phenotype data values and to at least one of the gene expression profile data values.

Databases useful in the invention can further include condition data values associated with the condition in the individual.

Also featured in the invention is a method of identifying a candidate mixture (e.g., a plant, fungal, bacterial, or animal tissue extract) for treating a condition in an individual, by (1) providing a first gene expression profile of a conditioned cell, the conditioned cell exhibiting a phenotype that can be correlated with the condition in the individual; (2) contacting the conditioned cell with the mixture having a plurality of (e.g., at least 10 or 100) different types of bioactive molecules; (3) determining a second gene expression profile of the conditioned cell after the contacting; and (4) comparing (e.g., by using a computer) the second gene expression profile with the first gene expression profile. A change in the first gene expression profile relative to the second gene expression profile indicates that the mixture is a candidate composition for treating the condition in the individual. A bioactive molecule is a molecule that elicits a biochemical or cellular response in

a cell. Bioactive molecules of the same type have the same
molecular structure.

The second gene expression profile can be determined by a
process including (1) isolating mRNA from the conditioned cell;
5   (2) producing cDNA (e.g., a labeled cDNA) from the mRNA; (3)
hybridizing the cDNA to an array of nucleic acid elements, each
element corresponding to a gene; and (4) measuring the extent to
which cDNA has bound to each element, thereby determining the
second gene expression profile. The array of nucleic acids can
10   be bound to a glass or silicon-based support.

The method of identifying a mixture can further include
(1) providing a third gene expression profile of a cell, the cell
being conditioned to produce the conditioned cell; and (2)
comparing the second gene expression profile with the third gene
15   expression profile. The greater the similarity between the
second gene expression profile and the third gene expression
profile, the greater the likelihood that the candidate mixture
can treat the condition in the individual. By "conditioned" is
meant any process which biologically, physiologically,
20   genetically, or biochemically alters at least one characteristic
of a cell. For example, exposing a normal cell to ionizing
radiation to induce immortal growth is a type of conditioning.
In this example, the alteration is likely to be genetic, as well
as physiological. Alternatively, the conditioned cell can
25   exhibit a normal phenotype, while the cell that is condition is
cancerous. This can be achieved by introduction of a vector
expressing an anti-oncogene (e.g., p53) into a cell to condition
it.

In addition, the invention includes a method of
30   identifying a candidate compound for treating a condition in an
individual by (1) providing a first gene expression profile of a
conditioned cell, the conditioned cell exhibiting a phenotype
that can be correlated with the condition in the individual; (2)
contacting the conditioned cell with a mixture; (3) determining a
35   second gene expression profile of the conditioned cell after the
contacting; and (4) comparing the second gene expression profile

- 4 -

with the first gene expression profile. A change in the first
gene expression profile relative to the second gene expression
profile indicates that the mixture contains the candidate
compound for treating the condition in the individual.

5   This method can further (1) providing a third gene
expression profile of a cell, the cell being conditioned to
produce the conditioned cell; and (2) comparing the second gene
expression profile with the third gene expression profile. The
greater the similarity between the second gene expression profile
10   and the third gene expression profile, the greater the likelihood
that the candidate compound can treat the condition in the
individual. Once these additional steps are performed, the
method can include (1) fractionating the mixture to obtain a
fraction; (2) contacting the conditioned cell with the fraction;
15   (3) determining a fourth gene expression profile of the
conditioned cell after the contacting; and (4) comparing the
fourth gene expression profile with the second gene expression
profile and the third gene expression profile. A fourth gene
expression profile that is more similar to the third gene
20   expression profile than the second gene expression profile
indicates that the fraction contains the candidate compound for
treating the condition in the individual.

        Alternatively, the method of identifying a candidate
compound can further include (1) fractionating the mixture to
25   obtain a fraction; (2) contacting the conditioned cell with the
fraction; (3) determining a third gene expression profile of the
conditioned cell after the contacting; and (4) comparing the
third gene expression profile with the first gene expression
profile and the second gene expression profile. A third gene
30   expression profile that is more dissimilar to the first gene
expression profile than the second gene expression profile
indicates that the fraction contains the candidate compound for
treating the condition in the individual.

        In general, each of the gene expression profile data
35   values or gene expression profiles above can include expression
levels for at least 10 genes (e.g., at least 100 or 1000 genes).

The upper limit of the number of genes represented in the profile is of course dependent on the estimated total number of genes in the genome of the organism studied. The human genome has been estimated to contain about 100,000 genes.

5   In one aspect, the methods of the invention can be used to test a preexisting purified compound, mixture of compounds, or extracts, each currently without any pharmaceutical use, for its ability to alter gene expression in a cell in a manner consistent with a condition or disease. No biological assays are required

10   for this analysis, though they can be included as confirmatory assays. This process, because it begins with a purified compound and looks for a disease to treat, is hereby termed "reverse drug discovery."

       In another aspect, the methods of the invention also

15   allow testing of complex mixtures, such as extracts, containing numerous molecules for efficacy against a condition or disease in a patient, again without any biological assays required. These methods can be used as a first screen of complex mixtures or to validate mixtures believed to have potential pharmaceutical

20   utility. Subsequently, the mixtures can be fractionated using standard techniques such as chromatography, and each of the fractions tested in the methods of the invention.

       Other features or advantages of the present invention will be apparent from the following drawings and detailed

25   description, and also from the claims.

                    Detailed Description of the Invention

       The methods and various computer-related aspects of the invention rely on the use of gene expression profiles or a

30   database thereof in finding a condition or disease that is treatable in an individual. In other words, known purified compounds, with or without known pharmaceutical uses, can be screened to determine possible use in treating any condition that can be represented in a gene expression profile.

35   The advent of DNA chips and other arrays has greatly accelerated and simplified the acquisition of gene expression

- 6 -

profiles from biological samples. Therefore, the use of DNA
arrays and associated tools can be used with the methods of the
invention. A relatively comprehensive discussion of arrays can
be found in Bowtell et al., supra, which is summarized below.

5   Bowtell et al. and references cited therein are hereby
incorporated into this document.

I.   Generation of RNA from Biological Samples to Be Profiled

        RNA can be isolated from almost any abundant biological
10 sample using standard protocols or commercially available kits.
However, care must be taken in scrupulously identifying and
harvesting the cells from which RNA is to be isolated. Mistaken
identification can result in corrupting both the specific
expression profile for that cell type and the gene expression
15 profile database to which the specific profile belongs. One
means of ensuring purity of the cell sample is to purchase or
order a tissue or cell sample from a depository, such as the
American Type Culture Collection.

        Although large numbers of archival samples are
20 available in many clinical departments, often the samples are
sub-optimal with respect to RNA integrity, fixation, or critical
patient information. The establishment of suitable tissue banks
is a logical adjunct to any in-depth RNA analysis of human
tissue; repositories must address issues of appropriate
25 collection and storage and also ensure that the samples are
accompanied by appropriate patient information, including
treatment, outcome, epidemiological and family history data. The
National Cancer Institute (NCI) coordinates a centralized tumor
bank for North American researchers (http://www-
30 chtn.ims.nci.nih.gov/). Commercial tissue banks, such as
LifeSpan BioSciences, also provide access to a wide variety of
human disease tissues (http://www.lsbio.com/).

        Diseased tissue generally contains a mixture of normal
tissue, inflammatory cells, necrotic tissue and, in cancer
35 samples, areas of different grade. Similarly, healthy tissue
also includes a range of cell types. All of these elements can

- 7 -

combine to produce a complex RNA expression profile.
Microdissection capability is thus critical for microarray
studies involving tissues and is also useful for associated
technologies such as comparative genomic hybridization. Current
5   protocols for fluorescent labelling of RNA demand large
quantities of RNA, which impedes the use of microdissected RNA on
GeneChip7 and glass slide arrays. Laser-based microdissection
offers a means of more rapidly obtaining pure material than
conventional techniques. The commercially available laser
10   capture microdissection microscope (http://www.arctur.com) is
thus a valuable adjunct to microarray studies. Strategies for
using limited material include PCR-amplification of total cDNA
before labelling, or the generation of ^{33}P-labelled nucleic acids
(or targets) for filters and glass slides, as these require
15   relatively small amounts of total RNA. Xenografts provide an in
vivo means of amplifying limited amounts of tumor cell material
and may reduce levels of contaminating non-neoplastic host
tissue, although they may fail to recapitulate the expression
pattern of the primary tumor.
20   The process of obtaining large amounts of RNA
from a homogeneous cell population is greatly simplified when
using continuous cell lines. It is important to remember that
most microarray analyses do not measure absolute levels of RNA
but rather compare RNA levels between two samples. Attention to
25   technical details such as density, pH and possible effects of
inducers used in conditional systems is critical.

II.   Making and Using Microarrays
       Once a suitable RNA preparation is obtained,
30   interrogation of that RNA to produce a gene expression profile
can be performed using microarrays.
       The complexity of the available arrays remains a major
issue and relates to the current state of identification of all
the genes in a given organism and the clone sets that house these
35   genes. The complete sequence of the human genome and that of
many model organisms are or will be available. Until that time,

however, the sequence of most genomes except those of
Saccharomyces cerevisiae, *Caenorhabditis elegans*, and some
unicellular microbes are incomplete.

Traditional approaches to gene discovery such as the
5   cloning of recessive or dominant mutations or the genes encoding
specific proteins have identified approximately 7,000 human
genes. In contrast, large-scale expressed sequence tag (EST)
sequencing has greatly accelerated the rate of gene discovery.
Initially promoted by Craig Venter and associates, EST projects
10   have spread into both the commercial and academic arenas. In
1991, Merck and Washington University established a collaboration
that ultimately fostered the deposition of 480,000 human EST
sequences into GenBank. That number has now grown to over one
million human ESTs, particularly through the efforts of
15   Washington University and members of the IMAGE consortium
(Integrated Analysis of Genomes and their Expression;
http://www-bio.llnl.gov/bbrp/image/image.html), and more recently, through
those of the Cancer Genome Anatomy Project (CGAP;
http://www.ncbi.nlm.nih.gov/UniGene/ gene_discovery.html). EST
20   sequences are deposited in dbEST
(http://www.ncbi.nlm.nih.gov/dbEST/index.html), a division of
GenBank, in which an automated process called UniGene compares
ESTs and assembles overlapping sequences into clusters in a
similar manner to shotgun sequencing projects
25   (http://www.ncbi.nlm.nih.gov/UniGene/index.html). Some ESTs
correspond with known genes, but the majority represent partially
sequenced novel genes. Ideally each cluster would correspond
with one gene, but as several non-overlapping clusters may exist
for large or low abundance genes, the number of clusters is
30   likely to exceed the number of separate genes from whose sequence
they are derived. Additionally, errors in alignment programs can
produce false clusters (over-clustering). Clone sets, comprising
a single representative of each cluster (usually the most 5'
clone), are sold by licensed vendors (http://www-
35   bio.llnl.gov/bbrp/ image/idistributors.html).

The large number of ESTs identified by the Institute for

Genomic Research (TIGR, http://www.tigr.org/) are now publically available. TIGR mouse, human, zebrafish, rat and plant clones can be viewed in their respective Gene Indices databases, where they have been assembled into tentative consensus sequences (and

5   can be considered the equivalent of UniGene clusters) by comparison with TIGR and GenBank databases. The American Type Culture Collection provides single human and mouse TIGR clones, including a limited number of clones with the complete open reading frame of known genes

10   (http://www.atcc.org/hilights/tasc2.html).

Genome Systems (GS; http://www.genomesystems.com/) and Research Genetics (RG; http://www.resgen.com/) are the two IMAGE clone vendors with the most developed clone sets. Both GS and RG have undertaken a process of clone validation through restreaking

15   (to isolate single cells) and resequencing, as the original UniGene sets had a significant discrepancy between actual and designated clone sequence and many IMAGE clones were mixed.

In addition to providing individual clones and clone sets, both companies sell filters on which clones or

20   purified DNAs have been arrayed at high density to provide targets for reverse-transcribed probes and supply some clone sets both as bacterial colonies in microtitre plates and as PCR products. The latter have the advantages of avoiding both the risk of T1 phage contamination and the need to isolate plasmids

25   for PCR, a step some labs feel is essential to obtain clean DNA for arraying purposes.

GS has been able to add to the human IMAGE clones via its access to additional human cDNA from Incyte (http://www.incyte.com), an organization that has perhaps

30   sequenced more human cDNA (3 million) than any other. Access to the Incyte LifeSeq database is currently limited to approximately 25 pharmaceutical partners (no academic institutions have subscribed). Of the human clones in LifeSeq, 2.3 million are Incyte-proprietary. By using the Incyte clone set, GS has

35   recently produced a sequence-verified set of 9,844 human clones that includes many known genes present in the UniGene set (about

5,000 of about 7,400). The GS resequenced human clones are also
free from low-level T1 contamination present in IMAGE clones,
which can be a serious problem (http://www-bio.llnl.gov/bbrp/
image/phage.html).

5   Both RG and GS clone sets also contain a large number of
ESTs. Whether those ESTs present are likely to be expressed in
your favorite tissue or cell line is hard to predict. Little or
no information is provided concerning the basis for selection of
ESTs; they appear to represent clones from a range of libraries
10   with no preselection on the basis of biological interest. That
situation is changing as the focus of the Washington University
human EST project shifts to CGAP clones and companies provide
clones that have interesting expression patterns.

          EST sequences of other organisms, such as mouse, rat,
15   *Drosophila melanogaster*, and *Arabidopsis Thaliana*, have
accumulated at different rates. A limitation of the mouse EST
project (http://genome.wustl.edu/est/mouse_esthmpg.html) is that
sequencing has been carried out from the 5' end of cDNA. As the
length of the 5' ends of cDNA is variable, the number of clusters
20   obtained is greater than if oligo(dT)-primed cDNA had been
sequenced from the 3' end. As a result of this, and because
fewer mouse cDNAs have been sequenced and clusters are smaller,
the proportion of novel genes in the current mouse clone sets is
substantially fewer than in the human sets. Both RG and GS are
25   performing 3? resequencing of a collection of mouse clones.
These clones have been selected because they either correspond to
known mouse genes or because they appear to be related to other
genes of interest, thus effectively collapsing some of the
earlier clustering. Celera (http://www.celera.com/), a
30   commercial offshoot of TIGR, is sequencing the *Drosophila* genome
as a prelude to their human genome sequencing project; the
Drosophila sequence also should be ready soon.

          Obtaining the entire genomic sequence of *S. cerevisiae*
allowed a near-complete set of genes to be generated by PCR,
35   which have been arrayed and analyzed. Although it will be more
difficult to identify coding sequences in more complex organisms,

- 11 -

the convergence of genome and EST sequencing projects in the near
future will ensure the identification of non-redundant clone sets
that encompass all genes for a variety of other species.

Filter arrays have the advantage of being relatively
5   affordable and needing no special equipment to use, although
potential users should be aware that large format phosphorimager
screens may be required with larger filters. Filters are also
useful for scarce RNA (for example from microdissected tissue),
as only approximately 50 ng of total RNA is required for a single
10 experiment (100 ig of RNA is typically required for a fluorescent
probe). The major disadvantage of filters is that comparison of
expression between two samples requires hybridization of each
sample to separate duplicate filters, or to a single filter that
must be stripped and hybridized sequentially. The sensitivity of
15 lysed colony filter arrays is reported to be limited to high- and
medium-abundance genes. In contrast, hybridization of
fluorescently labelled nucleic acids to slide arrays or gene
chips can detect low abundance genes, an important point as most
genes fall within this category.   Direct comparison of
20 GeneChip7, slide, and filter arrays is required, however, to
settle the considerable debate concerning the relative
sensitivity of filters hybridized with ^{33}P-labelled targets versus
GeneChip7 or slides hybridized with fluorescent targets.
Commercial filter arrays of clone sets are available from
25 Clontech, GS, and RG.

Important considerations in the choice of array include
whether the clones used to produce the arrays are restreaked and
sequence-verified, whether DNA or lysed colonies are arrayed, and
the number of known genes and ESTs. Clontech filters only
30 include known genes, preselected and grouped for their
involvement in specific processes such as apoptosis. Current GS
filters use lysed bacterial colonies, whereas purified DNA is
arrayed on both Clontech and RG filters. The lower complexity
and higher purity of arrayed DNA is thought to increase the
35 sensitivity of these filters. GS is expected to release
sequence-verified human arrays that include a large proportion of

- 12 -

known genes present in the UniGene set (approximately 5,000).

GeneChip7 arrays and commercially available glass slides are at the more sophisticated end of microarray analysis and are extremely suitable for use in the methods of the invention. At present, options include Affymetrix GeneChip7 arrays (http://www.affymetrix.com) and slide arrays from Incyte (which has recently acquired Synteni (http://www.synteni.com/)).

Incyte does not sell slide arrays as such but provides a service whereby samples applied to slides and the data returned. Molecular Dynamics and Clontech have recently announced that they will also provide slide arrays (http://www.mdyn.com/). Genometrix (www.genometrix.com) provides custom synthesis of large numbers of low complexity arrays (up to several hundred probes). Using a proprietary method for arraying oligonucleotides, they mass-produce slides at low cost (approximately $10/array for orders of 1,000-10,000 individual arrays) and are developing devices for high-throughput analysis of these arrays.

Hyseq (http://www.hyseq.com/) have developed a novel method where hybridization of DNA targets with all possible pentamer or heptamer oligonucleotides allows inference of sequence from the pattern of oligonucleotide hybridization. This strategy has been applied to measuring the abundance of individual cDNA in libraries from tissues of interest, thereby providing an estimate of individual gene expression. Hyseq offers this type of analysis in house.

The first glass slide arrays were produced in Dr. Pat Brown=s laboratory at Stanford University (http://cmgm.stanford.edu/pbrown/index.html) and from there the technology spread. Brown=s web site also contains detailed specifications for building an arrayer and associated software. Additional protocols and some hardware details are available at http://chroma.mbt.washington.edu/ mod_www/.

Several companies have produced arrayers for sale. Each is a relatively simple XYZ axis robot that can position the print head with a similar degree of precision. Critical determinants

when choosing among them are whether the machine has a proven
track record in the field, the technical support network
available, cost, ease of use, capacity, features such as bar code
reader, temperature control, plate stacker, microtitre plate lid
5   remover, and pen design.

Beecher Instruments was started by one of the engineers
who developed the robot used by the NHGRI; it now sells an
equivalent arrayer and reader, which has the advantage of having
been successfully field tested for more than two years. The
10   BioRobotics microGrid (http://www.biorobotics.co.uk/) combines
many features into a compact machine. Initially designed for
robotic gridding of clones from 96- or 384-well microtitre plates
onto filters, it can also replica-plate libraries and be upgraded
to print glass slides and filters and re-array bacterial
15   libraries (also known as cherry picking). Genomic Solutions
(www.genomicsolutions.com) produces a complete system of arrayer,
hybridization station, reader, and analysis software. Genetic
MicroSystems (http://www.geneticmicro.com/) also produce a
relatively affordable machine. Molecular Dynamics conducts a
20   Microarray Technology Access Program (MTAP), where participants
gain early access to microarray technology developed by Molecular
Dynamics and Amersham. Molecular Dynamics produces arrayers for
MTAP participants.

One of the most important factors affecting the
25   performance of the arrayer are the shape, reproduciblity, and
durability of the pens (also referred to as tips, pins, and
quills). Uneven pens deliver unequally during a print run and
tax the abilities of image analysis programs. Precision tips are
available from several suppliers, including Beecher Instruments,
30   Majer Precision Engineering (http://www.majerprecision.com), who
custom-machine high-precision pens from a range of materials, and
Telechem
International (http://www.wenet.net/~telechem/), who also offer
related microarray equipment and consumables.

35   Filters are hybridized with ^{33}P-labelled probes and signal
is detected using phosphorimager screens. Phosphorimager systems

- 14 -

are produced by Molecular Dynamics, Packard Instrument, and Fuji.
The Packard Cyclone instrument is relatively low in cost but
offers a high degree of resolution for array work. Analyses can
be carried out by eye, by sending a GIFF data file via the Web to
5   a company to be read, or using commercial or public domain
software (see below).

The Affymetrix fluorescence reader, produced by Hewlett
Packard, is currently customized for GeneChip7 arrays. Hewlett
Packard plans to build readers capable of reading both GeneChip7
10   and glass slides. General Scanning (http://www.genscan.com/)
released the ScanArray 3000, a compact scanning confocal laser,
and Beecher Instruments sells a reader based on the machines used
at NHGRI and the National Cancer Institute (NCI). Like the
arrayer, the Beecher reader is not supported by a service network
15   but its high degree of sensitivity has provided a benchmark for
other commercial readers. Molecular Dynamics have recently
released the Avalanche reader, which is based on one developed
during the MTAP program.

The above readers are laser confocal scanning devices,
20   except for the Genomic Solutions reader, which uses a CCD camera
and filter blocks, facilitating upgrades to reading different
fluorophors. Direct comparison would be very useful and can be
carried out in the context of the methods of the invention.

25   III.   Data Analysis
A typical array experiment generates thousands of data
points. Informatics can be categorized as either >tools= or
>analyzers=. Tools include software that operate arraying
devices and perform image analysis of data from readers,
30   databases to hold and link information, and software that link
data from individual clones to Web databases. Some involve
fairly straightforward software but are nevertheless quite
extensive. The Brown laboratory
has made available software for operating custom built arrayers
35   (http://cmgm.stanford.edu/pbrown/mguide/
software.html).

The quality of image analysis programs is crucial for accurate interpretation of signals for slide and filters. Dr. Yidong Chen (NHGRI) has developed a sophisticated image analysis program for slides and filters, deArray, that is available but

5   not supported (www.nhgri.nih.gov/DIR/LCG/ 15K/HTML/). Mark Boguski and colleagues have developed software that is capable of both analyzing microarray data and linking to databases such as Entrez and UniGene, and this can be downloaded from the web (www.nhgri.nih.gov/DIR/ LCG/15K/HTML/).

10   Commercial readers and arrayers provide software for data analysis: Synteni have developed a sophisticated program for analyzing microarray data (GemTools); RG sells the Pathways package to analyze their filters; and the Visage suite can be purchased from Genomic Solutions, separate from their hardware.

15   Silicon Genetics (http://www.sigenetics.com) provides the GeneSpring package for analyzing data from Affymetrix GeneChip7 and other microarray experiments.

RNA expression analysis represents only one parameter by which cells or tissues may be characterized. Depending on the

20   experiment, epidemiological or molecular pathological data, genomic changes (gains or losses) or sensitivity to drugs may be additional parameters that will influence the interpretation of microarray data. The ability to combine RNA and protein expression data to comprehensively profile both transcriptional

25   and post-transcriptional changes in cells and tissues is particularly appealing, although the number of proteins that can be profiled at this stage is substantially less than the number of genes. Although it is more difficult to identify proteins that are differentially expressed, techniques for rapid and

30   reproducible two-dimensional gel protein separation and mass spectrometry-based protein identification make high-throughput proteomics a highly desirable adjunct to microarray RNA expression analysis. Thus, the methods of the present invention can include proteomic assays.

35   Without further elaboration, it is believed that one skilled in the art can, based on the above disclosure and the

- 16 -

description below, utilize the present invention to its fullest
extent. The following detailed description is to be construed as
merely illustrative of how one skilled in the art can practice
the invention and is not limitative of the remainder of the
5   disclosure in any way. Any publications cited in this disclosure
are hereby incorporated by reference.

A pattern of gene expression is obtained from normal
cells, diseased cells, and compound-treated cells using DNA
chips. The gene expression pattern is a representation of the
10   state of the cell in response to the disease or to treatment.
Comparison of gene expression by healthy, diseased, and treated
cells will in principle reveal patterns of gene expression that
are diagnostic for therapeutic as opposed to pathological
effects. Once validated, such patterns can be used to screen
15   complex compound mixtures for molecules with desirable
properties. This approach does not depend on knowing the precise
molecular mechanism of a disease; rather, it identifies sets of
genes as diagnostic for a disease state, without requiring a
specific knowledge of the contribution of the genes to disease.
20   Since the compounds identified from screening the complex
mixtures are totally based on the change of expression pattern,
the pattern recognition can be accomplished via bioinformatic
analysis. Results from this analysis can then indicate the type
of disease to be treated. Therefore, this invention will be most
25   useful in discovering drug leads for treating various human
pathologies; from cardiovascular to autoimmune disorder; from
infectious disease to cancer.

After the completion of human genome sequencing, the
invention will further allow us to identify groups of genes that
30   may share common regulatory elements from a single given DNA chip
experiment. This invention can thus provide a revolutionary
approach towards novel lead drug discovery against so far unknown
groups of genes that may share a genuine disease-relevant common
elements.
35   As described within this document, the mixture or
composition can include a plant extract, microbial broth, or

- 17 -

chemical library. Fractionation is defined in two ways. If a
mixture is from natural extract the fractions can be obtained
from column chromatography. If a mixture is a synthetic chemical
library from combinatorial synthesis, the sub-libraries (also
5   named fractions for simplicity) will then be prepared using the
appropriate combinatorial synthesis. Diseased cells are defined
as a biopsy or a cell line, either taken from diseased tissue or
treated with a defined stimulus.

       The gene expression fingerprints of some selected drugs
10   on appropriate disease tissue or microorganisms with the DNA
chips are compiled in a relational database to provide basis for
pattern recognition of a given drug. A common characteristic
pattern of structurally similar drugs usually indicates that
these drugs may work through a similar mechanism to exert their
15   drug effect. This common pattern can be used to identify new
drug leads using the pattern recognition analysis. This will
allow discovery of new drug leads with new chemical structures.
This new lead may serve as a basis for further structural
modification for improved pharmacokinetics, including decreased
20   side-effects.

       For primary screening on DNA chips, a sample of the
disease cells is treated with mixtures for a specified time
period. After the treatment the mixture is washed away and the
cells are lysed. Fluorescently-tagged cDNA is prepared by
25   reverse transcription of mRNA from both experimental samples.
After hybridization with the DNA chips and appropriate washing
steps, images are analyzed via laser scanning. A controlled
sample without treatment is also conducted with the same
procedure. The housekeeping genes are included in the expression
30   system for the purpose of a quantitative measurement. The
fluorescence intensity ratios for the control vs. the test
samples are determined, and the changes in gene expression are
compared. Mixtures that produce a change of gene expression
profiles are selected for secondary screening.
35   For secondary screening, fractionation is performed on
the selected screening mixtures obtained from the primary

- 18 -

screening. Aliquots of the disease cells are treated separately with each fraction. Gene expression profiles of the fraction-treated cells are compared with those tabulated in the primary screen. Comparison of the expression data obtained from the primary and the secondary screens may require further screens with selected sub-fractions until active components are identified. Alternatively, gene expression-altering activity may rely on a complex fraction, in which case no further fractionation is possible without decreasing or eliminating the mixture's bioactivity.

At some points in the procedure (e.g., during the sub-fraction screening), the gene expression pattern may indicate an additive effect between different fractions. Further analysis using combination screens of fractions could then be employed to identify multiple agents that exhibit the combined additive effect.

The genomes of several microorganisms have been completely sequenced, allowing for comprehensive analysis of gene expression under different conditions. DNA chips representing the genome of a given microbe (e.g., a bacterium, fungus, or yeast) are used to probe the gene expression profile in different growth conditions: free-living, host-associated, or with or without drug treatment. Since many genes required for host infection are expressed preferentially by microbes in the process of infection itself, DNA chips allows rapid identification of potential drug targets. Lead compounds can be evaluated simultaneously for impact on gene expression by both the infecting microbe and the host, thereby allowing the identification of potential side effects in parallel with an assessment of therapeutic efficacy.

Procedures similar to the primary and secondary screens as described above can be applied to discover drug leads for treating microorganisms.

A reference set of gene expression patterns are assembled for a given microorganism in the free-living state, in association with the infecting host, and if available, in

- 19 -

infected hosts treated with known drugs. Comparison of these patterns will allow the identification of potential targets and assessment of mixture effects on these targets.

5   Labeled cDNA probes are prepared from microbes cultured in vitro or isolated from an infected host, in the presence or absence of the compound mixtures. Changes in gene expression profiles in response to a given mixture are recorded and the mixture itself is selected for further fractionation. Fractions are evaluated for their ability to generate changes in gene

10   expression. Fractions that fail to modify the pathogenic expression profile are discontinued, whereas fractions that reproduce changes in the profile are sub-fractionated. This process is repeated until candidate structures that are responsible for the change of the expression pattern can be

15   identified. In some cases, of course, the activity will be dependent on a mixture of compounds.

     Bioinformatic analysis is often used to decipher the compiled expression data with regard to the gene expression fingerprint library established by the above procedures. Two

20   scenarios are expected from the analysis.

Scenario 1:   Fractions or compounds that exhibit similar
              expression pattern to that of the known drug-treated
              cell may indicate that these fractions or compounds
              intervene the cellular process through a similar

25   biological mechanism. However, the discovered drug
              lead with new chemical structure may serve as a
              basis for further structural modification for
              improved pharmacokinetics, thereby decreasing the
              potential side-effects of the drug.

30 Scenario 2:   If a fraction or compound shows a completely
              different expression pattern than that from the
              known drug treated tissue sample, the fraction or
              compound may serve as a new drug lead for a novel
              drug target with a new mode of action in intervening

35   the disease process. This scenario will provide
              novel strategies in discovering drug leads for so

far unknown drug targets with a completely unknown
drug action.

The structures of the identified compounds are
characterized using conventional instrumentation analysis.
5 Structures that are amenable to organic synthesis will be
prepared and scaled up in the laboratory for both in vitro and in
vivo testing in an established disease tissue, disease animal
model, intact microorganism, or infected tissue. Once the lead
drug is validated, medicinal or combinatorial chemistry can be
10 applied to conduct optimization of the drug lead into a
clinically useful drug candidate.

As opposed to conventional drug development, the natural
extract or combinatorial libraries used as described in this
invention for drug lead discovery can provide another opportunity
15 to discover a desirable disease treatment method using a selected
combination therapy. This can be discovered through the
observation of the combined effect of different fractions on gene
expression patterns which are considered most desired in the
disease treatment.

20   There are two criteria that must be met by cell lines to
be useful in a microarray-based screening assay. First, the cell
lines should individually or as a set be able to model
physiologically normal and disease states. Secondly, it should
be possible to isolate sufficient mRNA to generate fluorescently-
25 labeled cDNA (approximately 10 ìg, which is typically isolated
from 5×10^6 cells).

An example of such an analysis for rheumatoid arthritis
(RA) was recently reported (Heller et al., Proc. Natl. Acad. Sci.
USA 94:2150, 1996). The human chondrosarcoma cell SW1353
30 releases matrix-degrading metalloproteinases (MMPs) when treated
with TNFá or IL-1. Since TNFá production contributes to joint
destruction in RA, the gene expression profile of TNFá-induced
SW1353 cells should be reflective of the disease state.
Comparison of the profiles from uninduced and TNFá-induced SW1353
35 cells could therefore generate a diagnostic gene expression
profile. TNFá-induced SW1353 cells would then be treated with

- 21 -

compound mixtures for screening purposes.

     Adherent cells are grown in flasks to generate a least 5
x 10^6 cells per flask. For each preparation of mRNA from TNFá-
induced cells treated with a mixture, there would need to be an
5 equivalent amount of mRNA from untreated TNFá-induced cells to
serve as a reference. The two sets of mRNA are used to generate
cDNA by reverse transcription. The reverse transcription
reactions contain dCTP labeled with either of fluorescent dyes
Cy3 or Cy5 (Amersham Pharmacia), which results in the generation
10 of fluorescently-labeled cDNA probes. Cy3 and Cy5 have similar
excitation spectra but distinct emission spectra. The two sets
of cDNAs are combined and hybridized to one or more DNA chips.
After washing to remove unbound probe, the chip is analyzed by
laser scanning. Since Cy3 and Cy5 have different emission
15 spectra, the amount of each probe hybridized to a given nucleic
acid (corresponding to a known gene) on the chip can be
quantitated separately, and the ratio of the two signals will
indicate whether expression of the gene has changed as a result
of exposure to the mixture. Several companies provide software
20 packages that allow for the compilation of gene expression
profiles from microarray data (e.g. GeneSight, from BioDiscovery,
Inc., Los Angeles, CA). For purposes of screening, any change in
mixture-treated cells compared to untreated cells is indicative
that the mixture can be used to treat a condition, but particular
25 emphasis will be given to mixtures that reverse the effect of
TNFá induction.

Other Embodiments

     It is to be understood that while the invention has been
30 described in conjunction with the detailed description thereof,
the foregoing description is intended to illustrate and not limit
the scope of the invention, which is defined by the scope of the
appended claims. Other aspects, advantages, and modifications
are within the scope of this invention.
35   For example, instead of determining a gene expression
profile associated with a condition or disease de novo, known

gene expression profiles can be used. For example, profiles for
multiple sclerosis lesions and corresponding normal tissue is
known (Whitney et al., Ann. Neurol. 46:425-428, 1999). In
addition, profiles for young and aged skeletal muscle are also
available (Lee et al., Science 285:1390-1393, 1999).

What is claimed is:

1.   A method of determining a candidate condition that is
treatable with a chemical composition, the method comprising

        providing a relational database comprising phenotype data
values and gene expression profile data values, wherein each of
5   the gene expression profile data values is related to at least
one phenotype data value, and each of the phenotype data values
is associated with a condition in an individual;

        contacting a cell with the chemical composition;

        obtaining a test gene expression profile of the cell
10 after the contacting;

        querying the relational database with the test gene
expression profile to obtain a test phenotype; and

        determining the candidate condition associated with the
test phenotype.
15

        2.   The method of claim 1, wherein the database further
comprises cell identity data values, each of the cell identity
data values being related to at least one of the phenotype data
values and to at least one of the gene expression profile data
20 values.

        3.   The method of claim 1, wherein the test gene
expression profile is obtained by a process comprising

        isolating mRNA from the cell;
25   producing cDNA from the mRNA;

        hybridizing the cDNA to an array of nucleic acid
elements, each element corresponding to a gene; and

        measuring the extent to which cDNA has bound to each
element, thereby determining the test gene expression profile.
30

        4.   The method of claim 3, wherein the cDNA is labeled.

        5.   The method of claim 3, wherein the array is bound to
a glass support.
35

6. The method of claim 3, wherein the array is bound to a surface of a silicon-based material.

7. The method of claim 1, wherein the chemical composition contains a pure compound or a mixture of a number of pure compounds.

8. The method of claim 1, wherein the chemical composition contains a plant extract.

9. The method of claim 1, wherein the chemical composition contains an extract of an animal tissue.

10. The method of claim 1, wherein each of the gene expression profile data values includes expression levels for at least 10 genes.

11. The method of claim 10, wherein each of the gene expression profile data values includes expression levels for at least 100 genes.

12. The method of claim 11, wherein each of the gene expression profile data values includes expression levels for at least 1000 genes.

13. A method of identifying a candidate mixture for treating a condition in an individual, the method comprising
      providing a first gene expression profile of a conditioned cell, wherein the conditioned cell exhibits a phenotype that can be correlated with the condition in the individual;

contacting the conditioned cell with a mixture comprising
a plurality of different types of bioactive molecules;

determining a second gene expression profile of the
conditioned cell after the contacting; and

5   comparing the second gene expression profile with the
first gene expression profile,

wherein a change in the first gene expression profile
relative to the second gene expression profile indicates that the
mixture is a candidate mixture for treating the condition in the
10   individual.

14.   The method of claim 13, wherein the mixture is a
plant extract.

15   15.   The method of claim 13, wherein the mixture is an
extract from an animal tissue.

16.   The method of claim 13, wherein the second gene
expression profile is determined by a process comprising
20   isolating mRNA from the conditioned cell;
producing cDNA from the mRNA;
hybridizing the cDNA to an array of nucleic acid
elements, each element corresponding to a gene; and
measuring the extent to which cDNA has bound to each
25   element, thereby determining the second gene expression profile.

17.   The method of claim 16, wherein the cDNA is labeled.

18.   The method of claim 16, wherein the array is bound
30   to a glass support.

19.   The method of claim 16, wherein the array is bound
to a surface of a silicon-based material.

35   20.   The method of claim 13, further comprising

- 26 -

providing a third gene expression profile of a cell, wherein the cell is conditioned to produce the conditioned cell; and

comparing the second gene expression profile with the third gene expression profile,

5   wherein the greater the similarity between the second gene expression profile and the third gene expression profile, the greater the likelihood that the candidate mixture can treat the condition in the individual.

10

21.   The method of claim 13, wherein the comparing step is computer-assisted.

22.   The method of claim 13, wherein each of the first
15   and second gene expression profiles includes expression levels for at least 10 genes.

23.   The method of claim 22, wherein the first and second gene expression profile each includes expression levels for at
20   least 100 genes.

24.   The method of claim 23, wherein the first and second gene expression profile each includes expression levels for at least 1000 genes.

25

25.   The method of claim 13, wherein the plurality is at least 10 different types of bioactive molecules.

26.   A computer system comprising

a memory storing a database comprising phenotype data values and gene expression profile data values, wherein each of the gene expression profile data values is related to at least one phenotype data value, and each of the phenotype data values is associated with a condition in an individual;

an input device configured to provide a test gene expression profile obtained from a cell after contacting the cell with a composition; and

a processor configured by a program to query the database using the test gene expression profile to obtain a test phenotype.

27. The computer system of claim 26, further comprising an output device for conveying the test phenotype or the condition associated with the test phenotype.

28. The computer system of claim 25, wherein the database further comprises cell identity data values, each of the cell identity data values being related to at least one of the phenotype data values and to at least one of the gene expression profile data values.

29. A computer-readable medium having a program adapted to configure a machine to query a database using a test gene expression profile to obtain a test phenotype, wherein the database comprises phenotype data values and gene expression profile data values, each of the gene expression profile data values relating to at least one phenotype data value, and each of the phenotype data values associated with a condition in an individual.

30. The computer-readable medium of claim 29, wherein the database further comprises cell identity data values, each of the cell identity data values being related to at least one of the phenotype data values and to at least one of the gene expression profile data values.

- 28 -

31.   The method of claim 1, wherein the relational database further comprises condition data values associated with the condition in the individual.

5

32.   The method of claim 2, wherein the relational database further comprises condition data values associated with the condition in the individual.

10

33.   The computer system of claim 26, wherein the database further comprises condition data values associated with the condition in the individual.

34.   The computer system of claim 28, wherein the

15   database further comprises condition data values associated with the condition in the individual.

35.   The computer-readable medium of claim 29, wherein the database further comprises condition data values associated

20   with the condition in the individual.

36.   The computer-readable medium of claim 30, wherein the database further comprises condition data values associated with the condition in the individual.

25

37.   A method of identifying a candidate compound for treating a condition in an individual, the method comprising

providing a first gene expression profile of a
conditioned cell, wherein the conditioned cell exhibits a
phenotype that can be correlated with the condition in the
individual;

5   contacting the conditioned cell with a mixture;

determining a second gene expression profile of the
conditioned cell after the contacting; and

comparing the second gene expression profile with the
first gene expression profile,

10   wherein a change in the first gene expression profile
relative to the second gene expression profile indicates that the
mixture contains the candidate compound for treating the
condition in the individual.

15   38.   The method of claim 37, further comprising

providing a third gene expression profile of a cell,
wherein the cell is conditioned to produce the conditioned cell;
and

comparing the second gene expression profile with the
20 third gene expression profile,

wherein the greater the similarity between the second
gene expression profile and the third gene expression profile,
the greater the likelihood that the candidate compound can treat
the condition in the individual.

25

39.   The method of claim 38, further comprising

fractionating the mixture to obtain a fraction;

contacting the conditioned cell with the fraction;

determining a fourth gene expression profile of the
conditioned cell after the contacting; and

5   comparing the fourth gene expression profile with the
second gene expression profile and the third gene expression
profile,

wherein a fourth gene expression profile that is more
similar to the third gene expression profile than the second gene

10   expression profile indicates that the fraction contains the
candidate compound for treating the condition in the individual.

40.   The method of claim 37, further comprising

fractionating the mixture to obtain a fraction;

15   contacting the conditioned cell with the fraction;

determining a third gene expression profile of the
conditioned cell after the contacting; and

comparing the third gene expression profile with the
first gene expression profile and the second gene expression

20   profile,

wherein a third gene expression profile that is more
dissimilar to the first gene expression profile than the second
gene expression profile indicates that the fraction contains the
candidate compound for treating the condition in the individual.

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :A01N 63/00; G01N 33/48

US CL : 424/93.7; 702/19

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 424/93.7; 702/19

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y,P	US 6,114,114 A (SEILHAMER et al.) 05 September 2000, see especially cols 7-21 and 77-82.	1-40
Y,P	WO 9957130 A1 (GENE LOGIC, INC.) 11 November 1999, see especially pages 1-35 and 49-55.	1-40
Y,P	WO 00/29846 A2 (CURAGEN CORPORATION) 25 May 2000, see especially pages 62-83.	1-25
Y	WO 99/10536 A1 (YALE UNIVERSITY) 04 March 1999, see especially pages 5-86 and 5-68.	1-25
Y	WO 99/01581 A1 (GENZYME CORPORATION) 14 January 1999, see especially pages 4-10 and 23-25.	1-25

[X] Further documents are listed in the continuation of Box C.   [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)		
		"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 NOVEMBER 2000	03 JAN 2001

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	SHUBO "JOE" ZHOU
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet) (July 1998)*

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US00/26050

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category°	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 99/10535 A1 (YALE UNIVERSITY) 03 March 1999, see especially pages 4-18 and 62-63.	1-25

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/26050

Box I   Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
   because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
   because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
   because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II   Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

   Please See Extra Sheet.

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest   ☐ The additional search fees were accompanied by the applicant's protest.
                          ☐ No protest accompanied the payment of additional search fees.

Form PCT/ISA/210 (continuation of first sheet(1)) (July 1998)*

B. FIELDS SEARCHED
Electronic data bases consulted (Name of data base and where practicable terms used):

DIALOG, MEDLINE, EAST

search terms: Huang, (gene expression profile), (disease or condition), databases

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION WAS LACKING
This ISA found multiple inventions as follows:

This application contains the following inventions or groups of inventions which are not so linked as to form a single inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Group I: Claims 1-12, and 26-36.

Group II: Claims 13-25, and 37-40.

Groups I and II are two distinct inventions because the invention of Group I is directed to methods and computer system/medium for determining a candidate condition mostly by looking for **similarities** between gene expression profile of cells contacted with certain chemical composition and that of any phenotype associated with certain condition in a relational database whereas invention of Group II is directed to methods of identifying a candidate mixture for treating a condition by looking for **differences** in gene expression profile of cells treated with certain mixture from that of a phenotype/condition in the database. Therefore, the Special Technical Feature is different between the above two Groups.

Form PCT/ISA/210 (extra sheet) (July 1998)*